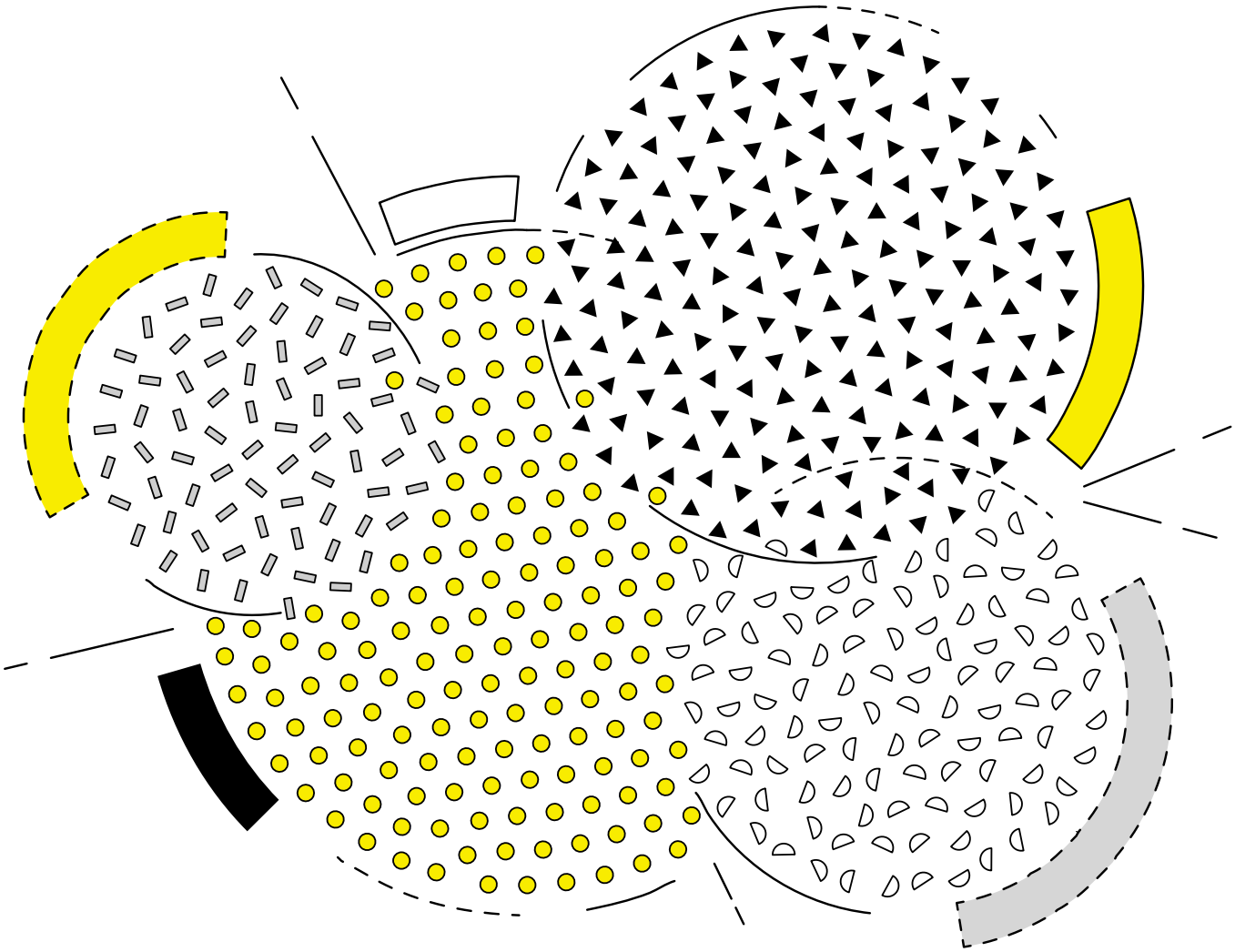
NOODL

AN EXPERIMENT IN EQUITABLE DATA LICENSING: PROMISE AND LIMITS.



OPEN _FUTURE

APRIL 2026

Alek Tarkowski

INTRODUCTION

This report analyses the Nwulite Obodo Open Data License (NOODL), a tiered licensing framework developed for African language datasets, as an experiment in open data licensing and a contribution to emerging approaches to data commons governance. The report has been commissioned by the Mozilla Foundation and supported by GIZ.

Over the past two decades, open licensing has been promoted as a means of ensuring broad access to knowledge and fostering innovation in the public interest. Yet open frameworks have failed to adequately account for—and address—the power imbalances related to sharing.

The NOODL license is a response to this challenge. Developed through consultation with African language communities, it builds on Creative Commons licensing but introduces a tiered framework of obligations based on geographic and economic context, with preferential terms for users in the Global South. Rather than treating all users identically, it acknowledges that meaningful openness requires differentiation based on capacity and position.

This report does not aim to evaluate how the NOODL license works within the specific African context for which it was designed—that assessment belongs to the communities and practitioners who use it. Rather, it examines NOODL as an innovation in open licensing that has relevance beyond its immediate context: for the broader debate on how to democratize AI, and for the various initiatives exploring equitable approaches to data sharing.

In its current form, NOODL remains an experiment—a single license applied to a single dataset. Its significance should not be measured by the number of its users. It matters because it opens a space for alternatives to traditional open licensing, and points to the need to move beyond binary debates of open vs closed systems. A healthy sharing ecosystem needs to support experiments like NOODL in order to remain relevant and resilient.

Thinking more broadly, there is an urgent need to build alternatives to the AI development stack optimized for extraction and efficiency. Trebor Scholz and Mark Esposito argue that we need a "solidarity stack" that is cooperatively owned and equitable.[1] The NOODL license fits well within this alternative AI development stack.

THE EQUITY GAP IN OPEN LICENSING

Over the past two decades, open licensing has become synonymous with democratization and public benefit. The promise was elegant: by removing restrictions, we enable broader access, collaboration, and innovation. Yet after two decades, there is a clear sense that there are growing challenges to equitable openness. Open Future framed the challenge in 2020 as the Paradox of Open. At the heart of this paradox lies an equity gap: open frameworks treat all users as formally equal, while the conditions under which they operate are profoundly unequal.

[1] Scholz, R. Trebor, and Marco Esposito. "Building a Solidarity Ecosystem for AI." Stanford Social Innovation Review, 2026. https://doi.org/10.48558/WP28-E151.

Open frameworks continue to challenge proprietary ownership, but they simultaneously enable concentrations of power, or at least are prone to capture for proprietary benefit, without value being given back to the commons. Without safeguards, transparency, and cultural shifts, openness disproportionately serves actors with greater economic resources and technological capacity, to the detriment of those who contribute to shared, open resources but lack such power.

The balance of gains and losses related to opening up resources is thus no longer clearly positive. Sarah Pearson from Creative Commons recently noted that "It is unlikely that the incentives to share can outweigh all of the growing countervailing forces at play: economic, moral, safety, more. We cannot respond by accepting these risks and harms as inherent and inevitable costs of public sharing knowledge".

In the free software world, similar dynamics prompted licensing interventions over the past two decades. The GNU Affero General Public License (AGPL) was developed in response to the realization that copyleft obligations do not apply to companies that run open source software without distributing it. Similarly, the "anti-Tivoization provisions" in the GPLv3 license addressed hardware manufacturers who used hardware locks to prevent users from running modified versions. Both interventions recognized that control over infrastructure or hardware can neutralize the benefits of openness, if that privileged position is abused.

The challenge is now recognized also in relation to content and data licensing. And the asymmetry behind this challenge has to do with underlying power imbalances tied to technical, economic, and geopolitical factors. Without acknowledging these imbalances, it becomes hard to prevent the extraction of value.

Standard open licenses treat all users as if they have the same capacity (economic, infrastructural, and skill-based) to benefit and to make use of the shared resources. A researcher in Nairobi accessing an openly licensed language dataset faces the same formal terms as a multinational technology company. However, this formal equality obscures a profound inequality. Corporations possess the computing infrastructure, capital, and talent necessary for rapidly developing commercial products. Thus, rather than democratization, we see the extraction of value, where communities contribute valuable data and well-resourced actors capture its value.

These power imbalances have a strong global dimension. As Ramya Chandrasekhar and Renata Ávila have recently argued, open resources created by Global South communities risk being absorbed into closed AI infrastructures, in a dynamic they term "commonswashing".[2] This dynamic becomes visible with regard to linguistic data, which is a key aspect of various societies' and communities' cultures and a valuable resource for training AI technologies.

As the role of Natural Language Processing (NLP) grows, researchers and language communities have discovered that they are being systematically positioned as data collectors rather than innovators. There is an urgent need to address neglect of African language data that results in

[2] Ramya Chandrasekhar and Renata Ávila, "What the India AI Summit Leader's Declaration Means for the Future of the Digital Commons," TechPolicy.Press, March 6, 2026, https://www.techpolicy.press/what-the-india-ai-summit-leaders-declaration-means-for-the-future-of-the-digital-commons/.

under-representation of these languages in new AI technologies.[3] And while the open source toolbox—including open data sharing frameworks—offers a tested solution, it risks disenfranchising local language communities from their linguistic resources.

Even worse, the speed with which commercial solutions are being developed means that they become available to users in Africa much more quickly than any local solutions that could be produced. This analysis, which provides a basis for designing new licensing solutions, adds nuance to the understanding of the Paradox of Open. A "level playing field" established through open licensing not only favors stronger players but also gives them a first-mover advantage. This means that their solutions can crowd out a market even before local actors can make use of the openly shared resources due to their limited resources.

Entities with a privileged position of power have the necessary resources (including compute power) to synthesize local datasets, combine with other data, enhance their value, and use them to build AI products or services. These offer superior capabilities to those built just on the locally sourced open data, available to local entities. Therefore, concentrations of power enable market monopolization and hinder the market placement and scale of local innovation.

This context is important for understanding the purpose of designing the Nwulite Obodo Open Data License (NOODL), as an intervention on behalf of African communities and languages. NOODL is a tiered license, building on Creative Commons licensing, with preferential terms for developers in the Global South.[4] It is a response to Global North companies extracting value from linguistic datasets created by African communities, without giving back to these communities.

In other words, it is a response to the current condition, in which traditional open licensing models turn out to be—at least in some contexts—inequitable. In the words of Dr. Chijioke Okorie and Dr. Melissa Omino, who designed the license: “Those in the latter group have access to and control of other resources and infrastructure needed for AI innovation; when combined with African datasets, this puts them at a considerable advantage. Market dominance and open access to datasets cannot be designed and applied as if all persons requiring access to datasets were similarly situated when they are not—that would be inequitable.”[5]

---

[3] Adebara, Ife. "AI and Language Data Flaring in Africa: Addressing the Low-Resource Challenge." Policy Brief No. 216. Centre for International Governance Innovation, November 5, 2025. https://www.cigionline.org/publications/ai-and-language-data-flaring-in-africa-addressing-the-low-resource-challenge/.

[4] Omino, Melissa. "The NOODL License: Licensing African Datasets to Support Research and AI in the Global South." InfoJustice, June 5, 2025. https://infojustice.org/archives/46434.

[5] Okorie, Chijioke I., and Melissa Omino. "Addressing Inequitable Openness in Licences for Sharing African Data and Datasets Through the Nwulite Obodo Open Data Licence." *Law, Technology and Humans* 7, no. 3 (2025). https://doi.org/10.5204/lthj.4001.

THE NOODL PROJECT AS A CASE OF AI DEMOCRATIZATION

The concentration of capacity in a handful of corporations, located predominantly in the United States and China, represents one of the defining power asymmetries for AI development.[6] The AI technology stack is heavily centralized and controlled by a few economic actors in a few geographic regions. And even if there are signs of "downstream" competition at the level of various applications and services, this diversity depends on a narrow set of inputs from an oligopoly offering models, cloud computing, or semiconductors.[7]

Machine learning technologies—the dominant training approach used in AI development today—exhibits traits typically associated with natural monopoly.[8] Competition analyses have identified several structural constraints: monopolistic tendencies in the infrastructure sector due to limited chips and suppliers, exclusive access to training data, high switching costs for cloud providers, and the "winner-takes-most" nature of general-purpose model training created by significant economies of scale.[9]

This concentration extends across the entire AI stack—from hardware and infrastructure to data, models, and applications. A small number of companies control the compute infrastructure and cloud infrastructure, but also the access to the data needed for AI training. The result is what researchers have termed the "compute divide": a dependency that affects smaller, academic research, and public interest initiatives alike.[10] Empirical research by Tamay Besiroglu shows that academic-only research teams are increasingly underrepresented in compute-intensive research areas, particularly foundation models. And this divide has a strong global dimension as well, and strongly impacts the Global Majority.

This concept can be extended to include a "data divide", as lack of training data, and thus representation of languages in LLMs is a key development challenge that determines the development of new models. The development of language models has been focused on a handful of languages, and the vast majority of languages do not have datasets on which to build

---

[6] United Nations. (2024). Governing AI for Humanity: Final Report of the High-Level Advisory Body on Artificial Intelligence. United Nations Publications.

[7] Von Thun, Max, and Daniel A. Hanley. "Stopping Big Tech from Becoming Big AI: A Roadmap for Using Competition Policy to Keep Artificial Intelligence Open for All." Mozilla Foundation, 2024. https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/Stopping-Big-Tech-from-Becoming-Big-AI.pdf.

[8] Narechania, T. N. (2022). "Machine Learning as Natural Monopoly." Iowa Law Review, 107(4), 1543.

[9] Sieker, Felix, Alek Tarkowski, Lea Gimpel, and Cailean Osborne. White Paper on Public AI: A Public Alternative to Private AI Dominance. Open Future / Bertelsmann Stiftung, n.d. https://openfuture.eu/publication/white-paper-on-public-ai/.

[10] Besiroglu, Tamay, Sage Andrus Bergerson, Amelia Michael, Lennart Heim, Xueyun Luo, and Neil Thompson. "The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?" Preprint, arXiv, January 4, 2024. https://doi.org/10.48550/arXiv.2401.02452.

language models.[11] As a result, most AI systems do not serve most of the world's linguistic communities, creating a new form of digital exclusion.[12] Therefore, access to high-quality and legally—and ethically—usable data, available as a public good, is a necessary condition for democratizing AI.

Democratizing AI, therefore, must mean more than making AI tools more widely available. It means restructuring the conditions under which AI is developed so that actors beyond these dominant corporations have the capacity to develop systems that serve the needs of their communities. Democratizing AI should also entail building capacity to understand AI, develop AI, use AI, distribute value generated by AI, and govern AI systems.[13] As the T20 Policy Brief on Democratizing AI for the Public Good argues, the global majority cannot realize the developmental potential of AI technologies without the capacity to develop local solutions, driven by the needs of their local communities. These solutions must create value for communities rather than extract it.[14]

This critical analysis of the political economy of AI development, focusing on the concentrations of power and existing inequalities in the pace of model development, is at the heart of the NOODL license initiative. It assumes that when datasets from the Global Majority are governed by frameworks designed in the Global North and optimized for actors with maximum computing resources, the result is that Global South communities contribute essential resources to systems they cannot build themselves, for purposes they do not control.

The new license is designed to shift this dynamic. The NOODL project, therefore, applies the idea of democratization of AI capabilities to the data layer and assumes that data licensing can offer leverage in support of such democratization.

RECONSIDERING OPENNESS

The NOODL initiative is not an isolated effort. It emerges at a moment when multiple actors are experimenting with novel approaches to data governance. Most ambitiously, Mozilla's Data Collective and Common Voice (which have supported the development of the NOODL license)

---

[11] Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. "Beyond English-Centric Multilingual Machine Translation." Preprint, arXiv, October 21, 2020. https://doi.org/10.48550/arXiv.2010.11125.

[12] Pava, Juan N., Thomas S. Mullaney, Caroline Meinhardt, Audrey Gao, and Diyi Yang. "How Can AI Support Language Digitization and Digital Inclusion?" White Paper. Stanford Institute for Human-Centered Artificial Intelligence (HAI), n.d. https://hai.stanford.edu/assets/files/hai-silicon-white-paper-how-can-ai-support-language-digitization.pdf.

[13] Seger, Elizabeth, Aviv Ovadya, Divya Siddarth, Ben Garfinkel, and Allan Dafoe. "Democratising AI: Multiple Meanings, Goals, and Methods." In AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 715–722. New York: ACM, 2023. https://doi.org/10.1145/3600211.3604693.

[14] Gimpel, Lea, Daniel Brumund, Alek Tarkowski, Maximilian Gahntz, Urvashi Aneja, Vukosi Marivate, and Anita Gurumurthy. "Democratizing AI for the Public Good: Key Concepts and Recommendations." Open Future / Digital Public Goods Alliance, n.d. https://openfuture.eu/publication/democratizing-ai-for-the-public-good/.

aim to enable responsible sharing of AI datasets while retaining control and ownership. In doing so, they works with the idea of introducing friction as a key design mechanism for commons-based sharing.

Already in 2018, the Responsible AI Licenses (RAIL) framework was proposed as a way of sharing AI assets while mitigating misuse through behavioral-use clauses.[15] Hugging Face has implemented mechanisms that allow data stewards to register access requests, monitor usage, and assess impact without fully closing off their datasets. Harvard Law School Library's Institutional Data Initiative (IDI) is experimenting with gated access to public domain datasets. Creative Commons has proposed CC Signals and a vision of reciprocity as a key principle for open sharing.

An important precursor is also the work of Te Hiku Media, a Māori organization that has been developing means of community control over language data. The Kaitiakitanga Licence established principles of data guardianship and reciprocal flow of benefits back to data communities. The creators of the license have assumed that "By simply open sourcing our data and knowledge, we further allow ourselves to be colonised digitally in the modern world."[16] This approach to data sovereignty is shared by the creators of NOODL.

What unites these initiatives is a recognition that the traditional approach to openness —the idea that ensuring as much access as possible produces the most social benefit—is rooted in a worldview that no longer is true. The baseline assumption has been that the resources necessary for benefiting from openness are widely and equally available. It is now clear that in contexts where that is not true—where actors lack, for example, computing infrastructure, capital, or access to markets necessary to build AI solutions—openness can become a social liability.

This shift represents what might be called "commons-based protectionism": the use of available licensing and governance mechanisms to protect shared resources from extraction. Paul Keller notes that these are pragmatic experiments by organizations fundamentally committed to open access, and should be read as endorsements of enclosure: "These bottom-up approaches are best understood as efforts to shore up the commons under adverse conditions, not to weaponize access or to recreate artificial scarcity as a governing principle".[17]

The frictionless nature of open licensing has been seen as one of its greatest strengths. Open licenses are indeed a form of sharing with very low friction. This is due to the fact that a broad range of permissions is given in advance, and to standardized licensing terms. NOODL licensing is based on a different assumption: that friction to sharing is a way of ensuring equitable access.

---

[15] McDuff, Daniel, Tim Korjakow, Scott Cambo, Jesse Josua Benjamin, Jenny Lee, Yacine Jernite, Carlos Muñoz Ferrandis, Aaron Gokaslan, Alek Tarkowski, Joseph Lindley, A. Feder Cooper, and Danish Contractor. "On the Standardization of Behavioral Use Clauses and Their Adoption for Responsible Licensing of AI." Preprint, arXiv, February 7, 2024. https://doi.org/10.48550/arXiv.2402.05979.

[16] Te Hiku Media. "Kaitiakitanga License." GitHub, n.d. https://github.com/TeHikuMedia/Kaitiakitanga-License/blob/tumu/LICENSE.md.

[17] Keller, Paul. "Abundance vs. Scarcity: Who Controls the Internet After AI?" TechPolicy.Press, January 22, 2026. https://www.techpolicy.press/abundance-vs-scarcity-who-controls-the-internet-after-ai/.

NOODL AS AN EXPERIMENT IN OPEN LICENSING

When Dr. Chijioke Okorie, Dr. Melissa Omino, and their colleagues —the team that developed the NOODL license—began articulating the need for an alternative to standard public open licenses, they encountered resistance from an unexpected quarter: advocates for openness itself. A license meant to ensure equity was seen as potentially enclosing the commons. It is important to refute this criticism, as NOODL is neither a proprietary nor a closed license. The tiered access model assumes that additional limitations are equitable and beneficial to the data community.

Popular open licensing frameworks, including Creative Commons licenses, offer a spectrum of licensing approaches, from an almost fully open CC0 mechanism to licenses that also include limitations on commercial use or prohibition of derivatives. Each license represents a different answer to the question posed by NOODL license creators: open to whom, and for what purposes? The fundamental insight behind the NOODL license is that formal equality produces substantive inequality in contexts of radical and growing power asymmetry. At the same time, NOODL includes key mechanisms of an open license. At its core, it is a standardized open license with attribution and share-alike components.

Also, the NOODL license should be understood in the context of prevalent open licensing approaches on the African continent. These are, to a large extent, shaped today by funding bodies which, according to Dr. Okorie, wield enormous power to support dataset creation, but also to shape rules on how they are shared. These organizations typically include a mandate to release data and content under CC0. In Okorie's views, these organizations should be less prescriptive, and various entities should have greater freedom to decide which license to use for a given dataset.[18]

The strength of this intervention is rooted in conversations—and deep understanding—with the language communities for which the NOODL licensing framework has been developed. In the words of Dr. Melissa Omino, "It's not that they're saying they don't want to share or they don't want to be open. They're saying that the way open is existing right now is actually harming them." Dr. Omino notes that, for a long time, there seemed to be a consensus that a universal concept of openness exists and can be translated into a universal approach to open licensing. NOODL starts with a different assumption: that open sharing needs to be based on the ability of a local community to control how their common resources are shared. This is due in part to the specific subject matter of the license—language datasets are cultural assets, communally owned, and with a history of preservation and curation despite injustice due to colonial power and control. It is this context that requires a revision of what open sharing means for these languages.

The fact that NOODL was developed in consultation with language communities that were made aware of how open licensing works is a significant shift. "I had an exercise where I asked workshop participants to think about a dataset creation project and consider which licence

[18] Okorie, Chijioke. "It's the NOODL License – Awesome and Amazingly Geeky!" SSRN, June 30, 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5339254.

made sense for them to use in sharing the created datasets. What stood out again was that many people in the NLP space were using open licences without fully understanding them. This has to change," writes Dr. Okorie.[19]

And indeed, over the years, there have been two clear trends in the open licensing space. First, there has been limited innovation around licensing frameworks. The CC0 mechanism, released in 2009, is the last such innovation. This lack of innovation is surprising, taking into account the changes to the digital ecosystem in the last 15 years. The second trend is the prioritization of the most open licenses, including CC-BY, CC-BY-SA, or CC0. The latter has been expressed through various institutional policies endorsed by funders, research bodies, governments, and by community norms in the digital commons movement.

The NOODL license is one of several initiatives that signal a shift from this approach. This shift is related, on one hand, to a growing awareness of harms related to AI development, and on the other to a need for tailoring varied, context-based solutions rather than "one size fits all" frameworks. In the words of Dr. Omino, openness needs to be based on willingness to understand a given, local context.

A TIERED LICENSING APPROACH

NOODL represents a shift in thinking from a "one size fits all" licensing model that has been at the heart of traditional open licensing frameworks. Rather than treating all users identically, it acknowledges that meaningful openness requires differentiation based on context and capacity.

The NOODL license is designed on the basis of a traditional open license, offering a standardized way to grant permissions, and providing legal clarity regarding rights over content (specifically, datasets). Specifically, it is an open license with Attribution and Share Alike conditions—the latter making the licensing model viral. At the same time, a tiered framework of obligations, based on a geographic-economic classification, is added to this standard framework. The users are differentiated based on their location, and the developmental status of the country where they are located. This approach acknowledges that formal equality in licensing terms produces inequitable outcomes when applied to actors with vastly different resources, infrastructure access, and market positions.

For users in Africa and other developing countries (defined as any country not classified as "high-income" by the World Bank), NOODL applies permissive Creative Commons-style terms. For users from developed countries NOODL introduced so-called "sharing of value or benefits". The creators of the license consider that benefit sharing does not need to entail only monetary compensation, but can also include infrastructure access, collaborative involvement in system development, or capacity-building support. The license leaves the specific value or benefit sharing mechanism open, by declaring only that the Data Recipient, if not located in a Developing Country, should transfer the benefit or value indicated by the Dataset Provider, prior to using the dataset.

---

[19] *Ibid.*

It is also important to note the limits of this differentiation as reflected in the legal text. The license does not include a general commercial-use restriction, which could be considered as the most typical form of addressing power inequalities, through differentiated licensing conditions. Rather, the differentiation operates through the imposition of benefit- or value-sharing obligations on users located in high-income countries. While NOODL is framed as an intervention against extractive dynamics, the enforceable provisions place the redistributive logic primarily in the benefit-sharing mechanism and its distinction between high-income and developing countries, rather than broader restrictions on commercial use.

SUPPORTING COMMUNITY DATA GOVERNANCE

Instead of a simple act of applying a standardized license, NOODL licensing requires community engagement to define the benefit-sharing requirements for a given dataset. These additional licensing mechanisms ensure that the sharing model expressed by the license expresses the needs and expectations of the given data community, considered as the steward of the language data.

Furthermore, for a broad category of potential users from the Global North, using a NOODL licensed dataset requires contact with the Data Providers in order to deliver the required benefit. From a traditional open sharing perspective, these would primarily be seen as barriers to sharing, generating unnecessary transaction costs. In the NOODL model, this is an opportunity to make the data community visible as a stakeholder in data reuse and AI development.

The first deployment of the NOODL license was tied to the development of DhoNam Dholuo Speech Dataset, a corpus for the Dholuo, one of Kenya's major indigenous languages spoken by 5 million people. The work on the dataset was led by Dr. Lilian Wanzare (at Maseno Center for Applied Artificial Intelligence), who has also been working on fair licensing practices for African language data. The process involved not just traditional data collection from community members, but also participatory data governance exercises. Through workshops, people involved in the dataset creation were involved in co-designing a data sharing framework for the DHoNam dataset—and they made the decision to adopt the NOODL license.

This shows that more complex data sharing mechanisms require a more robust governance practice: community engagement, and establishing relations that are missing in standard open frameworks. NOODL licenses cannot simply be picked, applied, and forgotten. They are envisioned as frameworks that are actively maintained, as benefits need to be defined, the tiered sharing framework needs to be managed, and as the value or benefit sharing process establishes a relationship between data producers and data users. They are also tools that enable community participation in data governance.

How NOODL operationalizes Data Commons governance

In order to better understand the NOODL license as an intervention in data governance, it can be understood as a licensing mechanism meant to support a data commons rather than just

ensuring open access and usage rights. A data commons framework also provides a lens through which the friction in open licensing can be better understood.

Commons-based approaches balance public interest, economic growth, and respect for fundamental rights. In other words, they offer a governance framework that balances data sharing with rules for protecting the interests of data subjects and creators and concerns over sustainability. There is today a growing interest in these data commons approaches, defined by the principle of "Sharing as much as possible, with as many restrictions as necessary"[20]. They can be understood as a result of a growing awareness of the paradox of open, and the inequities in open sharing.

In 2021, GovLab's Open Data Policy Lab argued for a new approach to Open Data that is "is concerned not simply with data itself but with the broader technical, social, political, and economic context within which data is produced and consumed."[21] It proposed an approach that shifts from an open default to purposeful publishing that prioritizes responsible use and data rights. Data commons is a term increasingly used to describe these solutions, based on collaborative governance of data in order to provide and ensure responsible access.[22]

Data commons in relation to AI technologies and training datasets have in recent years been explored by various organizations and initiatives, including also the Collective Intelligence Project,[23] CNRS and Open Knowledge Foundation,[24] or the Digital Public Goods Alliance.[25] In the African context, an "AI commons" approach to data sharing was proposed by the Open for Good initiative as a way for data communities in Africa to retain control over data, and avoid opportunistic exploitation that is the outcome of traditional open-sourcing of data.[26]

Commons-based approaches balance public interest, economic growth and respect for fundamental rights. In other words, they assume that innovation and responsible development can be coupled, if proper data governance methods are employed so that data sharing is balanced with rules for protecting the interests of data subjects and creators. At Open Future we

---

[20] Tarkowski, Alek, and Warso, Zuzanna. "Commons-based data set governance for AI." Open Future. https://openfuture.eu/publication/commons-based-data-set-governance-for-ai/.

[21] Verhulst, Stefaan, G., et al. "The Emergence of a Third Wave of Open Data." Open Data Policy Lab. https://opendatapolicylab.org/images/odpl/third-wave-of-opendata.pdf.

[22] Chafetz, Hannah, et al. "A Blueprint to unlock new data commons for AI". Open Data Policy Lab. https://incubator.opendatapolicylab.org/files/data-commons-for-ai-blueprint.pdf

[23] Huang, Saffron, and Divya Siddarth. "Generative AI and the Digital Commons." https://www.cip.org/research/generative-ai-digital-commons.

[24] Benhamou, Yaniv, and Mélanie Dulong de Rosnay. "Open Licensing and Data Trust for Personal and Non-Personal Data: A Blueprint to Support the Commons and Privacy." IIC – International Review of Intellectual Property and Competition Law 56 (2025): 1553–1589. https://doi.org/10.1007/s40319-025-01636-y.

[25] Nordhaug, Liv Marte. "The Role of Open Data in AI Systems as Digital Public Goods - Digital Public Goods Alliance." Digital Public Goods Alliance. https://www.digitalpublicgoods.net/blog/the-role-of-open-data-in-ai-systems-as-digital-public-goods/.

[26] Open for Good Alliance. "Open-Source AI Data Sharing: Yes! Data Colonialism: No!" *Medium*, October 31, 2023. https://medium.com/@openforgood/open-source-ai-data-sharing-yes-data-colonialism-no-3062a922de03.

seeked to conceptualize the data commons approach in more detail, and proposed a framework for commons-based data governance. Just like the NOODL license initiative, this framework assumes that a gradient of various data governance approaches is possible, rather than a "one size fits all approach". Open licenses are an important part of this gradient of solutions, but work only for some types of data, and in some contexts.

The framework offers six principles for governing datasets as a commons, based on which a range of mechanisms and frameworks can be developed. The six principles are: share as much as possible with necessary restrictions, ensure transparency, respect data subject choices, protect the commons, ensure quality, and establish trusted institution and community engagement.

NOODL operationalizes these principles specifically for datasets from the Global South, particularly African language and cultural datasets. The tiered use model based on users' geographic and economic position is an application of principle 1: Share as much as possible, with necessary restrictions. The principle recognizes that in contexts of power asymmetry, unconditional openness serves the powerful. NOODL maximizes sharing by making data openly available to all users, but with restrictions meant to ensure equity, through the benefit-sharing mechanism.

In addition to this, the NOODL license offers a strong application of principle 3: Respect the choices of data subjects and content creators. Language data, the primary focus of NOODL development, comes from speakers—individuals and communities who contribute their voices, their knowledge, their cultural expression. NOODL embeds respect for these choices in two ways: first, it mandates that communities be informed about how their data will be used and have the opportunity to consent; second, it gives communities power over subsequent commercial use through the benefit-sharing requirement. Communities are not merely data sources; they are stakeholders with legitimate say in how their contributions are used.

The creators of the NOODL license highlight the fact that the tiered obligations mechanism is also a way to establish connections with users of datasets. The aim is to offer a remedy to the current situation and a step towards a more equitable, long-term solution. Today, the linguistic communities are little more than an invisible source of raw data materials for AI development. The requirement of benefit sharing forces dataset users to connect with these communities and thus recognize them.

Finally, the tiered mechanism also supports principle 4: Protect the commons. This principle addresses sustainability: how do we ensure that shared resources remain generative and are not depleted through extraction? NOODL implements this principle in two ways. Benefit sharing from Global North users is meant to create flows of resources back to communities.

The NOODL framework also potentially includes mechanisms related to principle 6: Establish trusted institutions and ensure community engagement. Although this is not stated in the license, its creators envision a community decision-making process related to defining what kind of benefit sharing should be required from users in Developed Countries. Due to its complexity, successful deployment of the license also requires a trusted institution capable of managing access and ensuring that licensing terms are honored on behalf of the language communities.

The “African Dataset Creation Split Sheet”, an additional mechanism promoted with the NOODL license, fulfills principle 2: Be transparent about the data and provide documentation. The NOODL framework encourages dataset stewards to use the Split Sheet to provide transparent documentation of contributors, roles and percentage contributions. This data is meant to facilitate both attribution and benefit sharing.

Enforcement and Adoption Challenges

One of the key challenges to new licenses and other commons-based governance mechanisms are obstacles to their adoption and enforcement. In the case of NOODL, the key question is whether benefit-sharing can be enforced? If a company violates the license terms, what recourse do communities have? This question is not unique to NOODL, but it is particularly acute for licenses centered on benefit-sharing, which require monitoring and management of dataset usage, rather than one-time legal compliance.

It remains to be seen whether the NOODL license will see significant levels of use—for now it should mainly be seen as an intervention in the licensing ecosystem, without significant application yet—beyond the DhoNam dataset.

While open, standardized licensing frameworks scale relatively easily, more complex, commons-based sharing frameworks can be hindered by a combination of their complexity, lack of scaling mechanisms, and lack of familiarity among those looking for data licensing solutions. Furthermore, traditional open frameworks benefit from two decades of awareness raising work. CC0 and CC-BY, open licenses typically used and promoted as sharing frameworks in Africa, benefit from institutional familiarity.

A new license requires education, trust-building, and demonstration that it actually works. It remains to be seen how the NOODL project deals with this challenge. From its inception, creators of NOODL have not just drafted the legal license, but taken care to create a range of educational resources, and to conduct outreach with various stakeholders. They recognize that enforcement is not just a legal problem, and can also be solved through outreach and awareness raising. In addition, and different from a “one size fits all” approach, the NOODL license is not necessarily meant to scale.

Admittedly, the NOODL license in its current form does not offer full clarity with regard to the tiered access model. The value or benefit sharing framework is not fully developed—more resources and examples would help understand how it is meant to work in practice. For a community-driven project, there could also be stronger emphasis on ensuring that the language communities participate in this process—something that cannot be enforced by the license, but can be established as a norm. Finally, the tiered obligations for Developed Countries do not apply to China—where some of the major AI companies are located. One would expect that actors from the United States and China would be treated the same, taking into account the issue of concentrations of power in AI.

One of the key issues raised during the consultation of NOODL was the shape of tiered obligations. Early feedback raised questions about whether NOODL's developing/developed

country distinction was the right axis of differentiation. Some, particularly from the Lelapa.ai research lab, argued for an Africa/rest-of-world distinction instead, recognizing that there are also power asymmetries within the Global South. The Esethu license (developed by Lelapa.ai) is inspired by NOODL, but introduces a narrower geographic scope by distinguishing uses in Africa and in all other countries. Others suggested that sectoral differentiation might be necessary—health sector datasets might need different terms than language datasets. The existence of Esethu, a derivative license, shows both the need for licensing innovation and the accessibility of the NOODL framework.

CONCLUSION: SUPPORTING LOCAL INNOVATION IN OPEN LICENSING

The NOODL initiative is based on a vision of local, contextualized data governance, for which the frameworks and mechanisms are as homegrown as the community datasets that are being licensed. It provides a sovereign licensing approach that complements the development of local data commons and public AI solutions based on them.

NOODL emerges from a specific context—African language communities and researchers recognizing that standard open licenses were not serving their needs. But it speaks to a broader question facing data governance globally: how do we create frameworks that enable genuine democratization of AI, rather than just ensuring access?

Creators of NOODL are vocal about the need to build a broader spectrum of licensing approaches that move beyond the binary choice between open and closed solutions. Instead of a narrow set of standardized approaches, they offer a vision that accepts licensing heterogeneity in the name of equitable community stewardship of local data. They acknowledge challenges related to the proliferation of licensing options, but believe that the benefits of this approach outweigh the risks of a "one size fits all" approach. Similarly, they argue that licensing frameworks should not be fixed—they need to evolve with community needs or technological changes.

It remains to be seen how the NOODL license works in practice and whether it sees broader adoption. Yet already the very fact that it was proposed, in consultation with African communities and stakeholders in public AI development, is an important contribution to commons-based approaches to data governance.

Such experimentation should be embraced by the broader digital commons ecosystem and the various organizations that steward the commons. Authors of NOODL suggest the need to create an ecosystem for more diverse licensing approaches—for example, a platform where the various licenses could be shared, explained, and explored by legal experts, data communities, and AI developers alike. This work could be conducted by major data sharing platforms, such as Mozilla's Data Collective or Hugging Face.

There is also a need to support the broader deployment of these licensing models, in order to properly evaluate them, and further iterate their design. Greater complexity of commons-based

sharing—as compared to traditional open licensing—means that significant resources are needed to support these experiments, in all their phases. Deployment programs would also help understand to what extent localized solutions can be adapted to other communities and locales.

The NOODL license shows that introducing conditions and differentiation need not close off the commons. It further illustrates how licensing, far from being a purely technical matter, is fundamentally about governance: who has power, how that power is exercised, and whose interests are served.

ABOUT THIS REPORT

The work on this report has been commissioned by the Mozilla Foundation and supported by GIZ. Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the above-mentioned organizations.

Open Future is a European think tank that develops new approaches to an open internet that maximize societal benefits of shared data, knowledge and culture. Open Future advocates for Digital Commons—characterized by distributed production, collective governance, and shared stewardship—as offering the most viable path towards a resilient digital ecosystem.

Alek Tarkowski is the Strategy Director at Open Future. He holds a PhD in sociology from the Polish Academy of Science. He has over 15 years of experience with public interest advocacy, movement building and research into the intersection of society, culture and digital technologies. His current interests include public interest AI policies and AI dataset governance.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Chijioke Okorie, Dr. Melissa Omino, and Dr. Lilian Wanzare for their support in writing this report.